

Evaluating the quality of survey and administrative data with generalized multitrait-multimethod models*

DL Oberski¹ A Kirchner² S Eckman³ F Kreuter^{4,5,6}

¹ Tilburg University, The Netherlands

² University of Nebraska, United States

³ RTI International, United States

⁴ Institute for Employment Research, Germany

⁵ University of Mannheim, Germany

⁶ University of Maryland, United States

Abstract

Administrative register data are increasingly important in statistics, but, like other types of data, may contain measurement errors. To prevent such errors from invalidating analyses of scientific interest, it is therefore essential to estimate the extent of measurement errors in administrative data. Currently, however, most approaches to evaluate such errors involve either prohibitively expensive audits or comparison with a survey that is assumed perfect.

We introduce the “generalized multitrait-multimethod” (GMTMM) model, which can be seen as a general framework for evaluating the quality of admin-

*The authors are indebted to Hal Stern and Jörg Drechsler for their comments as well as Barbara Felderer for her assistance in preparing the data. This work was supported by the Netherlands Organization for Scientific Research (NWO) [Veni grant number 451-14-017].

istrative and survey data simultaneously. This framework allows both survey and register to contain random and systematic measurement errors. Moreover, it accommodates common features of administrative data such as discreteness, nonlinearity, and nonnormality, improving similar existing models. The use of the GMTMM model is demonstrated by application to linked survey-register data from the German Federal Employment Agency on income from and duration of employment, and a simulation study evaluates the estimates obtained.

KEY WORDS: Measurement error, Latent Variable Models, Official statistics, Register data, Reliability

1. INTRODUCTION

Register data and administrative records play an increasingly important role in statistics (Wallgren and Wallgren, 2007), and several authors recommend and predict the increased use of “big data” (Entwisle and Elias, 2013; Podesta, 2014), including administrative register data (Japiec et al., 2015). Uses to date include studies of how agricultural households affect land changes (Rindfuss et al., 2004), voter turnout (Ansolabehere and Hersh, 2012), or how peoples’ numerical ability relates to mortgage default (Gerardi et al., 2013). However, there is evidence that register data may contain considerable measurement errors (Groen, 2012). For example, Bakker (2012, p. 15) estimated that 24% of the variance in Dutch official hourly wages records was random measurement error; Ansolabehere and Hersh (2010, p. 1) reported that 16.1 million out of the 185.4 million listed voter registration records in the United States were invalid; and Ladouceur et al. (2007, p. 275) suggested that 20% to 30% of osteoarthritis cases are not registered in Quebec hospital administrative records, causing bias in prevalence estimates. The measurement error present in administrative records can severely bias and invalidate research results (Carroll et al., 2006; Saris and Gallhofer, 2007; Vermunt, 2010). It is therefore essential to evaluate the

extent of measurement error in register data.¹

The difficulty in studying error in register and administrative data, however, is that there is often no “gold standard” measure. Some authors have suggested to link administrative registers to a survey, assuming the survey contains no measurement error (e.g. Yucel and Zaslavsky, 2005). But measurement error in survey data is widespread (Hansen et al., 1961, 1964; Felligi, 1964; Andrews, 1984; Alwin, 2007; Saris and Gallhofer, 2007; Biemer, 2011), and is in fact often measured by taking administrative records as the “gold standard” (e.g. Kapteyn and Ypma, 2007; Kreuter et al., 2010; Sakshaug et al., 2010; Kim and Tamborini, 2014). Thus, we often have two data sources, both measured with error, and we are interested in estimating the error in both.

Very few studies have attempted to estimate measurement error in both survey and administrative data simultaneously. Nordberg et al. (2004) discussed a longitudinal latent Markov model of measurement error in income, but again assumed the administrative register to be perfect in cross-sectional data; Pavlopoulos and Vermunt (2013) applied a similar latent Markov model to unemployment data; and Bakker (2012) and Scholtus et al. (2015) estimated measurement error using linear factor analysis. However, the models used in these studies have several drawbacks when applied to administrative register data. First, true values of the variables of interest are often censored, zero-inflated, gamma, count, or nominal, and thus models which assume normally distributed true values are not appropriate. For example, income is usually zero-inflated and occupation is nominal. Second, the measurement error process in registers is likely to lead to nonnormal and nonlinear errors, yet many models used to study measurement error assume linear and homoskedastic errors. For example, top-coding of income causes nonlinear method effects (Gottschalk and

¹We use the terms “register data” and “administrative data” synonymously to avoid repetition.

Huynh, 2010), and it is often thought that low earners over-report while high earners under-report, yielding “mean-reverting” random errors (e.g. Kim and Tamborini, 2014). Third, the measurement quality of administrative data often differs over observations, yielding a mixture of measurement models. For example, the records may be obtained from a mixture of sources (Wallgren and Wallgren, 2007), such as both employer statements and employee self-reports, or the variable may be more ambiguously defined for some cases than for others: the income of day laborers is an example. Earlier approaches have not accounted for such heterogeneity. Currently, then, there is no generally applicable method to evaluate the extent of measurement error in register and survey data.

Our contributions to the literature are twofold: we present a framework for simultaneously estimating measurement error in register and survey data which addresses the shortcomings of earlier methods; we then provide guidance on the circumstances in which survey data or register data are preferable for use in research. Section 2 introduces the modeling framework used to estimate the extent of measurement error in survey and register data simultaneously, and demonstrates how this framework encompasses existing methods. Section 3 applies the model to linked survey-register data on income and duration of employment from the German Federal Employment agency, while a simulation study in Section 4 evaluates the estimates obtained.

2. MEASUREMENT ERROR ESTIMATION FROM MULTIPLE ERROR-PRONE SOURCES

Our technique for simultaneously estimating measurement error in survey and administrative data builds on the “multitrait-multimethod” (MTMM) approach (Campbell and Fiske, 1959). Given a set of variables of interest (“traits”) for which observed measurements exist in both the administrative data and a sample survey, our goal is

to estimate the degree of measurement error in variables observed in both sources.

Let y_{tm} denote an observed random variable measuring the t -th trait using the m -th method. In the application described here, m will denote either administrative or the survey measurement.

Example 1. Suppose true income from full-time jobs η_1 , part-time jobs η_2 , and other types of jobs η_3 are of interest, for instance for future study of their relationship with educational attainment. Corresponding error-prone observed measures y_{11} , y_{21} , and y_{31} are obtained in an administrative register. For a random subsample of cases, we also have survey measures of the same variables: y_{12} , y_{22} , and y_{32} . There are thus three “traits” (full-time, part-time, and other income) and two “methods” (register and survey), and six observed variables. An equivalent view is that y_{tm} results from a repeated measures design in which the factors “trait” and “method” have been fully crossed.

2.1 Current approaches to modeling MTMM data

Commonly, MTMM data are analyzed using the linear model

$$y_{tm} = \tau_{tm} + \lambda_{tm}\eta_t + \gamma_{tm}\xi_m + \epsilon_{tm}, \quad (1)$$

where τ_{tm} is the constant systematic bias in y_{tm} and λ_{tm} and γ_{tm} are constant scaling factors with respect to the random variables. The “trait factor” η_m is a random subject \times trait interaction, symbolizing the “true value” of the trait measured by y_{tm} . The “method factor” ξ_t is a random subject \times method interaction, symbolizing method bias that differs over subjects but is common to variables measured with the same method. The residual ϵ_{tm} is random measurement error.

Assuming all η_t , ξ_m and ϵ_{tm} follow a multivariate Gaussian distribution, Model 1 is a confirmatory factor analysis (CFA) model with parameter vector $\boldsymbol{\theta} := (\boldsymbol{\tau}', \boldsymbol{\lambda}', \boldsymbol{\gamma}', \boldsymbol{\sigma}'_{\eta}, \boldsymbol{\sigma}'_{\xi}, \boldsymbol{\sigma}'_{\epsilon})'$,

where the parameters have been collected into vectors and $\sigma_{\mathbf{x}}$ denotes the nonredundant elements of the covariance matrix of \mathbf{x} , stacked columnwise.

Under this model the implied product-moment correlation between two observed variables y_{tm} and $y_{t'm'}$ (for $t \neq t'$) is

$$\text{cor}(y_{tm}, y_{t'm'}) = \begin{cases} \lambda_{tm}^* \lambda_{t'm'}^* \text{cor}(\eta_t, \eta_{t'}) & \text{if } m \neq m' \\ \lambda_{tm}^* \lambda_{t'm}^* \text{cor}(\eta_t, \eta_{t'}) + \gamma_{tm}^* \gamma_{t'm}^* & \text{Otherwise,} \end{cases}$$

where $\lambda_{tm}^* = \lambda_{tm}[\sigma_{\eta_t}(\boldsymbol{\theta})/\sigma_{y_{tm}}(\boldsymbol{\theta})]^{1/2} = \text{cor}(y_{tm}, \eta_t)$ is the “reliability coefficient” of y_{tm} and $\gamma_{tm}^* = \gamma_{tm}[\sigma_{\xi_t}(\boldsymbol{\theta})/\sigma_{y_{tm}}(\boldsymbol{\theta})]^{1/2} = \text{cor}(y_{tm}, \xi_m)$ is the “method effect”. Thus, when the measures have been obtained by *different* methods, the correlation between two observed error-prone variables is attenuated by a factor $\lambda_{tm}^* \lambda_{t'm'}^*$ relative to the correlation between the “true scores” η_t and $\eta_{t'}$: a classical result (e.g. Lord and Novick, 1968; Fuller, 1987). This result shows that it is essential to model both random measurement error ϵ_{tm} and individual method biases ξ_m : their presence will have dramatically different effects on subsequent analyses of interest. The MTMM design allows for the separation of these two error factors.

This approach has led to a large literature on MTMM modeling using CFA (structural equation modeling) to estimate the degree of random and systematic measurement error in survey data (e.g. Alwin, 1973; Andrews, 1984; Saris and Andrews, 1991; Saris and Gallhofer, 2007; Bakker, 2012). Extension for ordinal categorical data using the “ordinal factor analysis” model (Muthén, 1983) have also been applied (Oberski et al., 2008). Recently, Oberski (2013) introduced a latent class factor (Vermunt and Magidson, 2004) MTMM model.

The MTMM framework is in principle attractive for the modeling of measurement errors in administrative and survey data. For register data, however, these

currently available MTMM models are inadequate and can yield biased or nonsensical estimates, for the three reasons given in Section 1: nonnormality of true values, nonlinearity and heteroskedasticity of errors, and the existence of unknown groups that exhibit differential measurement error. We generalize the MTMM framework to allow for these possibilities.

2.2 The generalized multitrait-multimethod model

We use generalized latent variable models (Skrondal and Rabe-Hesketh, 2004) to formulate a measurement model for MTMM data from an administrative register and a survey that can account for non-classical error processes, nonnormal distributions, and categorical data. Generalized latent variable models are built up from (i.) linear GLM predictors; (ii.) GLM links and exponential family distributions; and (iii.) conditional independence relations. The conditional independence relations we use result from the MTMM design and are common to all MTMM models, whereas the choice of links and distributions is flexible: for this reason we call our approach a “generalized multitrait-multimethod” (GMTMM) model. The flexibility in links allows us to model nonlinearities and heteroskedasticities in the error process, while the choice of distributions for the latent variables allows for nonnormality of the true values. Finally, when heterogeneous measurement error processes need to be accounted for, a finite mixture is used that allows the parameters of the linear predictors to differ over the mixture components.

(i.) Linear predictors. For continuous observed data, linear predictors for the observed variables y_{tm} are:

$$\nu_{tm} = \tau_{tm} + \lambda_{tm}\eta_t + \gamma_{tm}\xi_m, \quad (2)$$

where, for identification purposes, the first loading of each trait factor η_t and method factor ξ_m is often set to unity, $\lambda_{t1} = \gamma_{1m} = 1$. For categorical observed data, linear predictors for category $y_{tm} = k$ are

$$\nu_{ktm} = \tau_{ktm} + \lambda_{ktm}\eta_t + \gamma_{km}\xi_m, \quad (3)$$

where the first category can be chosen as a reference by setting $\tau_{1tm} = \lambda_{1tm}^{(\eta)} = \lambda_{1m}^{(\xi)} = 0$ (e.g. Vermunt and Magidson, 2013).

The above linear predictors are common to all population units, and therefore assume that the measurement process is homogeneous. When the error process is thought to be heterogeneous, the linear predictor parameters are allowed differ over the mixture components, yielding an additional subscript $\nu_{tm,s}$ or (for categorical data) $\nu_{ktm,s}$.

(ii.) Links and distributions. Each of the observed and latent variables is assigned a distributional “family” and a link function $g(\cdot)$ connecting the linear predictor to the expectation of the response y_{tm} is chosen,

$$g(E[y_{tm}|\eta_t, \xi_m]) = \nu_{tm}, \quad \text{or} \quad g(E[y_{ktm}|\eta_t, \xi_m]) = \nu_{ktm}, \quad (4)$$

depending on whether the observed variable is continuous or categorical.

We denote the choice of the conditional distribution of the observed responses given the latent variables as $f_y := p(y_{tm}|\eta_t, \xi_m)$ with parameter vector $\boldsymbol{\theta}_y$. Similarly, the multivariate distribution of the latent “true score” variables is denoted f_η with parameters $\boldsymbol{\theta}_\eta$ and the distribution of the latent “method” variables f_ξ with parameters $\boldsymbol{\theta}_\xi$. Depending on whether the variables to which they refer are continuous or categorical, f_y , f_ξ and f_η may be probability density or probability mass functions.

Finally, the finite mixture components are assigned a multinomial distribution.

(iii.) Conditional independencies. The specification of the generalized latent variable model is completed with assumptions of conditional independence that are necessary for identification of the model parameters from observables. These assumptions mirror those of the linear MTMM model.

Assumption 1. The observed variable y_{tm} is conditionally independent of all other observed variables given its trait factor η_t and method factor ξ_m .

Assumption 1 implies that the joint conditional distribution of observed given latent variables can be factored into the univariate conditional distributions, i.e.

$$p(\mathbf{y}|\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{\theta}) = \prod_{t,m} f_y(y_{tm}|\eta_t, \xi_m, \boldsymbol{\theta}_y). \quad (5)$$

Assumption 2. The latent method factors $\boldsymbol{\xi}$ are mutually independent and independent of the trait variables $\boldsymbol{\eta}$.

Assumption 2 implies that the latent variable joint distribution can be factored into

$$p(\boldsymbol{\xi}, \boldsymbol{\eta}|\boldsymbol{\theta}) = f_{\eta}(\boldsymbol{\eta}|\boldsymbol{\theta}_{\eta}) \prod_m f_{\xi}(\xi_m|\boldsymbol{\theta}_{\xi}). \quad (6)$$

Note that there may still be dependencies among the latent trait variables in the vector $\boldsymbol{\eta}$.

Example 1 (continued). The conditional independencies can be displayed in a graph with directed arrows for GLM regressions and undirected edges denoting possible (conditional) dependence. Figure 1 shows the GMTMM model for the six-variable MTMM data from Example 1. In the Figure, observed variables y_{tm} are shown as rectangles while unobserved random variables (factors) are shown as el-

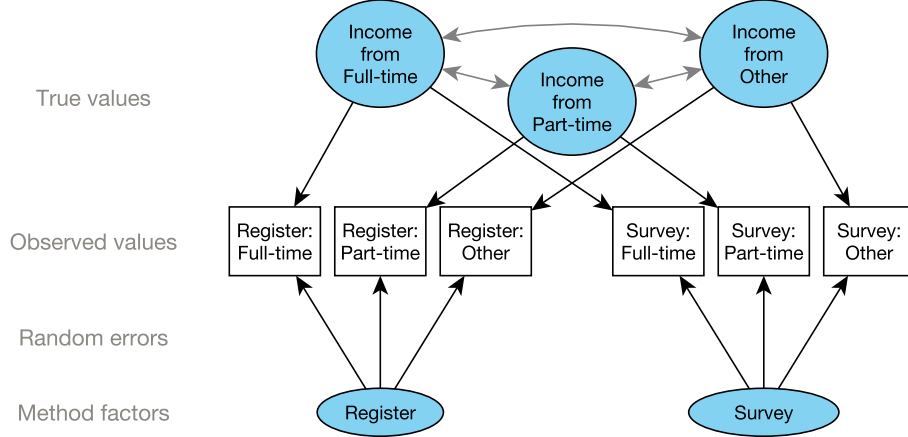


Figure 1: A generalized multitrait-multimethod (GMTMM) model for three “traits” using administrative data and a survey as measurement “methods”. The example traits signify personal income from full-time, part-time, and other kinds of employment over a certain period.

lipse. Assumption 1 can be verified by noting that conditioning on the hidden nodes yields an independence graph (e.g. Lauritzen, 1996).

Likelihood. When the error process is thought to be homogeneous, the marginal likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int \int \left[f_{\eta}(\boldsymbol{\eta}|\boldsymbol{\theta}_{\eta}) \prod_m f_{\xi}(\xi_m|\boldsymbol{\theta}_{\xi}) \prod_{t,m} f_y(y_{tm}|\eta_t, \xi_m, \boldsymbol{\theta}_y) \right] d\boldsymbol{\eta} d\boldsymbol{\xi}. \quad (7)$$

where assumptions 1 and 2 are used and the integral is defined as a sum for discrete latent variable distributions.

For heterogeneous error processes, in which a mixture of error processes is thought to be present, define $p(\mathbf{y}|S, \boldsymbol{\theta}_s)$ as the component-specific marginal likelihood, with component specific parameters $\boldsymbol{\theta}_s$. Typically, it is the measurement parameters that are thought to differ over components: that is, the linear predictors $\nu_{tm,s}$. We then introduce an unobserved discrete variable S with categories equal to the number of

components, so that the marginal likelihood of the observed data becomes

$$p(\mathbf{y}|\boldsymbol{\theta}) = \sum_S p(S)p(\mathbf{y}|S, \boldsymbol{\theta}_s). \quad (8)$$

Since the mixture proportions $p(S)$ are typically unknown, this implies an additional $|S|$ parameters in $\boldsymbol{\theta}$ to be estimated.

2.3 Special cases of the GMTMM model

By choosing different link functions, distributions, and error structures, a range of models that has been introduced in the literature to estimate measurement error in MTMM designs and administrative register data result as special cases of the GMTMM model.

Example 2. A common choice is to assume homogeneous errors, the identity link function $g(x) = x$, and distributions $f_y = N(\nu_{tm}, \sigma_{\epsilon_{tm}})$, with Gaussian latent variables $f_\eta = \text{MVN}[\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_\eta)]$, $f_\xi = N(0, \sigma_{\xi_m})$, leaving $\boldsymbol{\Sigma}(\boldsymbol{\theta}_\eta)$ unrestricted so that $\boldsymbol{\theta}_\eta = \boldsymbol{\sigma}_\eta$. This is the linear confirmatory factor analysis MTMM model presented above. This model was applied to linked survey-register data by Bakker (2012) and Scholtus and Bakker (2013).

Example 3. Leaving f_ξ and f_η unchanged from Example 2, the probit factor model for binary data results from choosing $f_y = \text{Binomial}[E(y_{tm})]$ with $g = \Phi^{-1}(\nu_{tm})$, where Φ is the standard normal distribution function. If, instead, the link function $g = \text{logit}(\nu_{tm})$ is chosen, a “two-parameter logistic” item response theory MTMM model is obtained.

Ordered categorical data can be modeled by choosing $f_y = \text{Multinomial}[E(y_{tm} = k)]$, redefining the observables, and choosing the link function

$$g[\text{Pr}(y_{tm} \leq k|\eta_t, \xi_m)] = \Phi^{-1}(\nu_{ktm}),$$

where the loadings are set equal over categories, $\lambda_{ktm} = \lambda_{tm}$, $\gamma_{ktm} = \gamma_{tm}$, and the category-specific intercept $-\tau_{ktm}$ plays the role of a cumulative probit “threshold” (Rabe-Hesketh et al., 2004). An ordered probit relationship between y_{tm} and the latent variables is thus specified. This model is known as the “ordinal factor analysis” model in the structural equation modeling literature (Muthén, 1983) and is a multidimensional version of the “normal ogive graded response model” in the item response theory literature (Samejima, 1969).

Example 4. The CFA and categorical CFA models in Examples 2 and 3 relied on normally distributed latent variables. It is possible to relax this assumption of normally distributed latent variables by specifying $f_\eta = \text{Multinomial}(\boldsymbol{\pi}_\eta)$ with free joint probability vector $\boldsymbol{\theta}_\eta = \boldsymbol{\pi}_\eta$, and univariate distributions $f_\xi = \text{Multinomial}(\boldsymbol{\pi}_{\xi_m})$, with free univariate probability vectors $\boldsymbol{\theta}_\xi = \{\boldsymbol{\pi}_{\xi_m}\}$. The number of latent categories to which f_η and f_ξ refer must be chosen in advance, yielding a finite mixture or “latent class” MTMM model (Oberski, 2013). When accompanied by the choice $f_y = \text{Multinomial}$, this model was described as “nonparametric” by Skrondal and Rabe-Hesketh (2004, sec. 4.4.2) and as “semiparametric” by Heinen (1996).

2.4 Estimation and identification of GMTMM model

The parameters $\boldsymbol{\theta}$ can be estimated from linked survey-register data when there are at least three “traits”—that is, variables of interest that have been measured with error in both survey and administrative register. Standard estimation procedures for generalized latent variable models can be used to estimate the GMTMM model (e.g. Skrondal and Rabe-Hesketh, 2004, chapter 6). The most general is to use standard optimization algorithms to maximize the marginal likelihood from Equation 7 or 8. For certain models, such as latent class MTMM models, direct maximization of the marginal likelihood may become unstable. An expectation-maximization (EM)

algorithm can then be used (McLachlan and Krishnan, 2007).

Many of the special cases of GMTMM models, including the examples given above, can be estimated using standard software for latent variable modeling such as Latent Gold (Vermunt and Magidson, 2013) or GLAMM (Rabe-Hesketh et al., 2004), that implement this estimation strategy. Moreover, specialized efficient estimation procedures already exist for certain special cases of the GMTMM model. For example, the linear factor analysis MTMM model can be formulated as a covariance structure model with a closed-form marginal likelihood (Bollen, 1989). The ordinal factor analysis (cumulative probit) model can be similarly dealt with by first computing polychoric correlation coefficients (Muthén, 1983). Such models can be fit using standard software for structural equation modeling. Other possible combinations of choices may require specialized software.

2.5 Model identification

The GMTMM model is a latent variable model, and its parameters are therefore not necessarily identifiable. A first point of interest is whether a given GMTMM model, such as the ordinal CFA MTMM model (Example 3), will have identifiable parameters. A second point of interest is what number of traits and methods are minimally required to identify the parameters of any GMTMM model. Assessing identifiability can be particularly relevant in advance of designing a survey to evaluate administrative data quality, since this will determine how many questions should be asked in the survey.

We take parameters to be “identifiable” if and only if a finite number of parameter values will lead to any given likelihood for all parameter values of nonzero measure (see Allman et al., 2009, for some of the subtleties involved in this definition). Trivially, for example, with only one variable observed on one trait using a

single method, it is clearly not possible to establish the parameter values regarding the latent trait and latent method factor variables separately, since infinitely many choices of $\boldsymbol{\theta}$ will lead to the same likelihood. On the other hand, the well-known “label switching” phenomenon in latent class-type models (McLachlan and Peel, 2000) leads to finitely many solutions and is therefore not considered an identification problem here. Similarly, choices of $\boldsymbol{\theta}$ that lead to rank deficiencies but have a point mass in the parameter space (see for example Shapiro and Browne, 1983) are not considered identification problems in this definition.

First, under the definition given, a given GMTMM model’s parameters will be identifiable if and only if the Jacobian $\partial p(\mathbf{y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is of full column rank almost everywhere (Catchpole and Morgan, 1997, Theorem 1). Equivalently, the rank of the information matrix may be examined. For GMTMM models with a closed-form marginal likelihood, this condition can be established analytically by assessing this rank using a symbolic algebra program. This may be considered an inconvenience by many applied researchers, however. For models without a closed-form marginal likelihood, analytical proofs are even more difficult. Numerical methods are then the more convenient tool to assess identifiability.

A common numerical approach is to examine the rank of the information matrix at the maximum likelihood estimate for a given dataset using the same software used to fit the model. The disadvantage of this method is that it conditions on the data at hand. For example, a model may appear identified when it is not, due to boundary solutions, and it may appear non-identified for particular parameter values when it is identified in the larger parameter space. To overcome this disadvantage, Forcina (2008) suggested evaluating the rank of the Jacobian at a large number of random values in the parameter space. This method has been implemented in the software Latent Gold 5 (Vermunt and Magidson, 2013).

This Section introduced a generalized multitrait-multimethod model that can be used to estimate measurement error when at least two separate measures of at least three different phenomena are available. The GMTMM model can deal with nonnormality of true values, nonlinearity and heteroskedasticity of errors, and the existence of unknown groups that exhibit differential measurement error. It is therefore applicable to estimating measurement error in administrative register data and surveys simultaneously. It is also more generally applicable to situations where such error structures are thought to exist in multiple error-prone sources.

3. APPLICATION TO ADMINISTRATIVE DATA ON INCOME AND DURATION OF EMPLOYMENT

This Section applies the GMTMM model to a unique dataset provided by the Institute for Employment Research (*Institut für Arbeitsmarkt- und Berufsforschung*, IAB), the research institute of the German Federal Employment Agency (*Bundesagentur für Arbeit*, BA). The BA's normal operations include job placement and payment of benefits, and for these purposes it maintains an extensive database of citizens' (un)employment histories dating back to 1975. This database covers German employees who are subject to social security contributions as well as recipients of entitlements, comprising about 86% of the overall German labor force. Excluded from the register are most civil servants, the self-employed, and others who have never been in contact with the Agency, such as the never-employed.

Both survey data and the BA's register data are routinely used for labor market and policy research—especially those on income and duration of employment. For consenting respondents, we gained IRB approval to link administrative record data from the Agency with a telephone survey conducted by the IAB (IAB Beschäftigtenhistorik (BEH) Version 09.01.00, Nürnberg 2012). Restricted access to the anonymized linked

survey-administrative data was provided at the Agency’s offices; the raw data cannot be made publicly available for legal reasons.

Particularly of interest are the BA’s records on *income* from full-time, part-time, and “marginal” employment. “Marginal” employment, also known as “Minijobs”, is a common form of low-income employment in Germany, yielding monthly income of up to 400 Euro (at the time of data collection); at or below this maximum, the employee is exempt from income taxes and social security. Of additional policy interest are the *durations* of the last employment spell of these three employment types. These data are not provided by the employees themselves, but rather by their employers, who are legally required to report their employees’ income accurately for the purposes of taxes, benefits, and social security.

However, exactly because the income and duration data were collected for the BA’s administrative purposes, measurement error can become a serious issue for research in spite of reporting accuracy, because measurement errors in administrative data need not come from the reporting itself (Bakker, 2009; Groen, 2012). For example, although the employers will presumably fulfill their mandate to report accurately, when compiling historical records there may be mismatches and time lapses in an individual’s record. Similarly, smaller jobs may simply be absent from the records, again leading to a mismatch in “last part-time job”, for instance. These issues will lead to random and correlated measurement error for research purposes.

To obtain the survey measurement, a stratified sample of 2,400 respondents was asked to provide information on income and employment duration from full-time, part-time, and marginal employment (see Eckman et al., 2014, for further description of the sample design). The survey had a response rate (AAPOR RR1) of 19.4%. In the following analyses, we accounted for the sample stratification using complex sampling adjustments. Of the respondents, 2,284 (95%) provided informed consent

to record linkage between the survey and the administrative registers. This linkage could be performed using unique person identifiers, so that it seems reasonable to assume no linkage errors were present. By linking the administrative data to the survey data, we thus obtained MTMM designs with three traits and two methods, one for each of the income and duration data.

The register provides income data only at the level of employment spells. This typically corresponds to an annual basis if a respondent was employed at the same employer throughout a given year. The survey, however, explicitly asks for the last monthly income from gainful employment which is the standard reference period used in most German surveys. Assuming that salaries are paid evenly throughout the employment spell, the administrative data were converted to a monthly basis.

3.1 Estimates of reliability and method effects in survey and administrative measures

To demonstrate the flexibility of the GMTMM approach and account for possibly differing measurement processes in the two measures investigated, we fit different types of GMTMM models to the duration and income data.

Duration data. For the duration data, we estimate Gaussian GMTMM models: that is, the familiar linear structural equation model using the standard SEM software `lavaan` for R (Rosseel et al., 2013; R Core Team, 2014). The program code to estimate this model can be found in the Appendix.

This approach yielded estimates for the trait loadings (λ_{tm}), method loadings (γ_{tm}), factor (co)variances ($\sigma_{\xi m}$, $\sigma_{\eta t}$), and error variances (σ_{tm}). In a linear model, the quality of each administrative variable can be simply represented by two numbers: the reliability and the method effect. These represent, respectively, the correlation between the observed administrative variable and its measured trait, and between

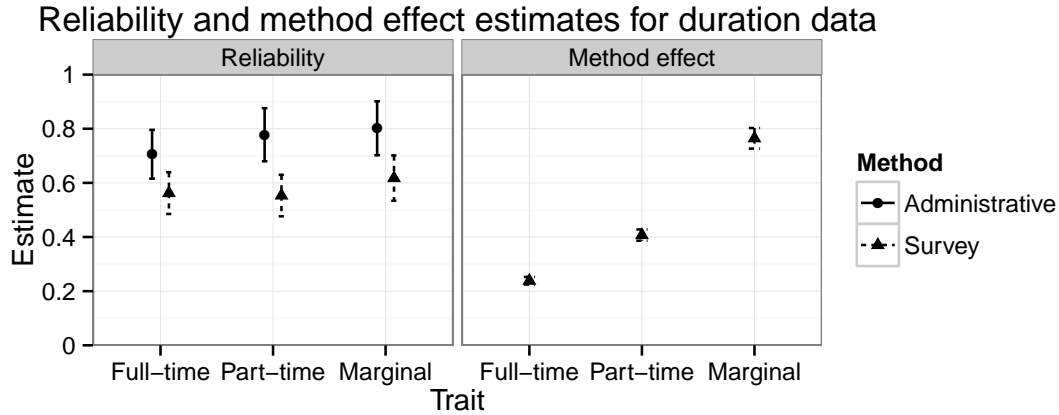


Figure 2: Reliability and method effect estimates for survey data, and reliability estimates for administrative register data on duration of full-time, part-time, and “marginal” employment.

the observed variable and the method factor (Saris and Gallhofer, 2007). A high reliability indicates that a survey question or register value contains little random error and accurately reflects the true value it measures. A high method effect, on the other hand, indicates that a substantial part of the variance is due to factors shared with other survey or register measures, but which are independent of the true values. An ideal measure would therefore have reliability one and zero method effect. Estimates of the reliability and method effects are displayed for the duration data in Figure 2.

Figure 2 shows reliability estimates in the left-hand panel and method effect estimates in the right-hand panel for the administrative and survey data on duration. The reliability estimates in Figure 2 are between 0.7 and 0.8 for the administrative data, which indicates that reliability of the administrative data is acceptable, but far from perfect. For example, the correlation between administrative records on full-time duration and the person’s true full-time duration is estimated at 0.7. The administrative measures’ reliabilities are clearly higher than the survey measures’

reliabilities, which are around 0.6. Thus, the self-reports were somewhat less reliable than the administrative records, but neither measure was perfect.

While fitting the model, the method effects (γ_{tm}) and method factor variances ($\sigma_{\xi m}$) for the administrative measures were estimated at zero but caused serious dependencies among the parameter estimates. We followed Eid (2000) and Saris and Gallhofer (2007) in fixing these to zero and re-estimating the model without method dependencies in the administrative data. The right-hand panel of Figure 2 therefore shows method effects for the survey measures only. These method effects can be seen as small for full-time durations, medium for the part-time durations, and very large for durations of “marginal” jobs. For example, a standardized method effect of 0.4 implies that answers to two survey questions on income will correlate by 0.4 above and beyond any true correlation between the two measures, thereby inflating relationship estimates that do not account for method effects. These large dependencies may be related to survey respondents’ different but systematic interpretations of a “duration”, or of what counts as a “marginal” job. However, there does not appear to be any such effect in the administrative data.

Income data. To estimate the quality of the administrative register as well as the survey answers on income data, we adapt the model to recognize several aspects of the measurement process:

- Following the econometrics literature (Tobin, 1958), censoring in income is accounted for;
- The relationship between true income and reported income is thought to be nonlinear (Kim and Tamborini, 2014);
- Previous studies linking survey and register data (Scholtus, 2015) suggested

that there is a subgroup of respondents for whom the two measures correspond exactly, whereas for others they do not, possibly suggesting a heterogeneous error process;

- There is a strong incentive to misreport one’s income from a “Minijob” as being equal to or below 400 euros, since at the time of the survey this was the legal maximum income to qualify for tax exemption and social security exemption (see §8 SGB [Social Security Code]).

Due to these factors, a linear Gaussian MTMM will not suffice. Instead, we choose f_y to be the standard censored regression equation, use the “nonparametric” latent class factor analysis formulation of f_ξ and f_η to allow for nonlinearity (Oberski, 2013), and investigate whether an additional mixture component of S in which the response is unrelated to the true value fits the data more closely than a homogeneous error structure. This model is no longer a standard structural equation model but can be estimated in the software for latent class (factor) analysis Latent GOLD 5.0 (Vermunt and Magidson, 2013). Program input can be found in the Appendix.

The latent class factor analysis model does not impose a distribution on the latent trait and method factors, but instead approximates these distributions by discrete interval-level latent variables whose category sizes are estimated from the data (Vermunt and Magidson, 2004). Moreover, the possibility of a heterogeneous error structure suggests the presence of an additional discrete nominal latent variable S . Since the number of categories for the latent trait, method, and error structure variables is unknown, we compare the fit of models with differing numbers of categories for each of these. Since increasing the number of categories of the method factors and the error structure variables beyond two never improved the model, we only show these comparisons for models with differing numbers of categories K for the

K	Error process							
	Heterogeneous				Homogeneous			
	LL	BIC	AIC	# par.	LL	BIC	AIC	# par.
2	-5060.0	10413.8	10195.9	38	-5388.3	11024.0	10840.6	32
3	-4758.3	9825.9	9596.6	40	-5272.1	10814.8	10614.1	35
4	-4848.9	10030.3	9783.8	43	-5210.1	10714.1	10496.3	38

Table 1: Fit of GMTMM models for the measurement error in administrative and survey data on income.

latent trait variables (η_t), with ($|S| = 2$) and without ($|S| = 1$) a heterogeneous error structure.

Table 1 shows the fit of these models in terms of loglikelihood (LL), BIC, and AIC, as well as the number of parameters these models have. The model with three latent categories and a heterogeneous error process fit the data best in terms of BIC and AIC. This result suggests that there may indeed be differing error processes for different respondents. Since the model fit did not improve when increasing the number of latent categories from three to four, we selected the three-class heterogeneous model. In other words, we approximate the distribution of true latent income with a discrete three-category latent variable for which the category sizes are estimated. We also allowed for some proportion of the observations to be unrelated to the true value, for example because some fixed value (such as 400 euros) was always chosen in this group regardless of the true income.

Table 2 shows the expected means of the administrative and survey measures of log-income for different categories of the latent trait and method variables. The table illustrates how the observed measures are estimated by the model to relate to the respective latent variables. The relationships in Table 2 are marginalized over the two categories of the error process latent variables S . Thus, the table shows how the relationship holds for a respondent whose error process is not known in advance.

	Trait			Method		Overall
	1	2	3	1	2	
<i>Administrative data (log-income)</i>						
Full-time	1.11	2.69	4.31			1.85
Part-time	0.65	1.54	2.45			1.08
Marginal	0.09	0.23	0.36			0.21
<i>Survey data (log-income)</i>						
Full-time	2.20	3.16	4.12	5.52	2.25	2.65
Part-time	0.91	1.67	2.45	1.44	1.26	1.28
Marginal	0.27	0.33	0.38	0.33	0.32	0.32

Table 2: Estimated relationships between categories of the latent trait variables η and the expected observation of log-income from full-time, part-time, and marginal employment using the administrative and survey measures.

About 5% (not shown in the table) are estimated to belong to the latent category in which a random value is given – that is, a value that is unrelated to the trait or method variables.

The model is no longer linear, so that reliability and method effect coefficients, which represent (linear) correlations are more difficult to interpret. However, it is possible to calculate the model-implied reliabilities $\text{cor}(y_{tm}, \eta_t)$ and method effects $\text{cor}(y_{tm}, \eta_m)$. These estimates, with confidence intervals based on bootstrapped standard errors, are shown in Figure 3. The figure shows that while the administrative data on income from full-time and marginal jobs are estimated to be superior to the survey measures, the survey measure has a stronger linear correlation with true income level from part-time work. A possible explanation for this difference is a change in mandatory reporting procedures regarding part-time employment in the year 2011. On the other hand, the survey measures do exhibit a strong method dependence, whereas again the administrative register measures were estimated to have no such method dependence.

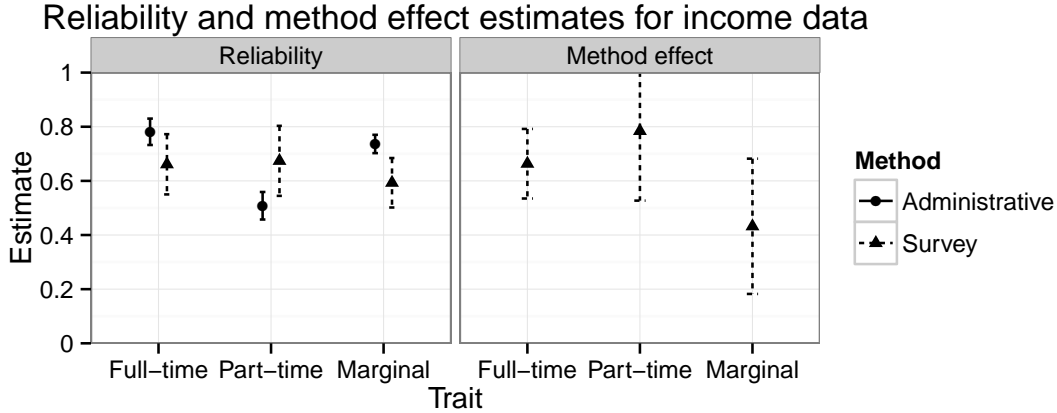


Figure 3: Reliability and method effect estimates for survey data, and reliability estimates for administrative register data on income from full-time, part-time, and “marginal” employment.

In summary, we found for official administrative data obtained from the German Federal Employment Agency that the reliability of both survey *and* administrative data was far from perfect. Estimated relationships between these observed variables and other variables of scientific interest will therefore be biased. Moreover, for some of these measures, method effects were found that will cause spurious dependencies where none exist among the true variables; when using administrative data, method dependence may be less of a concern. To prevent biases arising from measurement error in substantive analyses of income or duration data, correction methods for known error processes may be needed (e.g. Saris and Gallhofer, 2007; Vermunt, 2010; Skrondal and Kuha, 2012).

4. SIMULATION

We demonstrate some key properties of the maximum likelihood estimates of GMTMM model parameter estimates using a simulation study. Since there are many possible GMTMM models that fall within this framework, we choose the model and parameter values based on the linked survey-register dataset obtained from the German

Federal Employment Agency, and summarize bias and standard error accuracy under different conditions corresponding to sample sizes.

The response model chosen for the observed variables is a censored regression in which the unobserved trait and method variables are the regressors and the dependent variables are six observed indicators corresponding to the crossing of three traits and two methods. Thus, the response model for the observed variable y_{tm} measuring trait t with method m is

$$y_{tm} = \begin{cases} 0, & \text{if } y_{tm}^* \leq 0 \\ y_{tm}^*, & \text{otherwise} \end{cases}, \quad (9)$$

where y_{tm}^* follows the linear factor model,

$$y_{tm}^* = \tau_{tm} + \lambda_{tm}\eta_t + \gamma_{tm}\xi_m + \epsilon_{tm}, \quad \epsilon_{tm} \sim N(0, \sigma_{\epsilon,tm}). \quad (10)$$

The latent variables themselves are discrete interval-level variables with a multinomial distribution parameterized using the log-linear model

$$P(\eta_1 = k_1, \eta_2 = k_2, \eta_3 = k_3) = \frac{\exp(\mu_{k_1 k_2 k_3})}{\sum_{k'_1 k'_2 k'_3} \exp(\mu_{k'_1 k'_2 k'_3})}, \quad (11)$$

$$P(\xi_m = k) = \frac{\exp(\kappa_{mk})}{\sum_{k'} \exp(\kappa_{mk'})} \quad (12)$$

where $\mu_{k_1 k_2 k_3} = \sum_{t=1}^3 \alpha_{tk_t} + \phi_{12}\eta_{1,k_1}\eta_{2,k_2} + \phi_{13}\eta_{1,k_1}\eta_{3,k_3} + \phi_{23}\eta_{2,k_2}\eta_{3,k_3}$.

This model yields the following set of parameters, corresponding to the observed variable intercepts τ_{tm} , trait loadings λ_{tm} , method loadings γ_{tm} , error variances $\sigma_{\epsilon,tm}$, as well as the latent variable loglinear intercepts α_{tk} , and κ_{tk} and latent loglinear

associations $\phi_{tt'}$:

$$\boldsymbol{\theta} = (\{\alpha_{tm}\}, \{\kappa_{mk}\}, \{\tau_{tm}\}, \{\lambda_{tm}\}, \{\gamma_{tm}\}, \{\sigma_{\epsilon,tm}\}, \{\phi_{tt'}\})'$$

Furthermore, corresponding to the selected model from our application, we choose three categories for the latent trait and two for the latent method variables:

$$|\eta_t| = 3, |\xi_m| = 2.$$

To ensure parameter values are realistic, we set them to the maximum-likelihood estimates found in our application, and vary the sample size across conditions, $n \in \{200, 500, 1000, 2000\}$. The results of simulating data from this model and analyzing them using the GMTMM model are summarized in Table 3.

Table 3 summarizes the bias, defined as the difference between the true parameter value and the simulation average of the maximum likelihood estimate, as well as the ratio between and the ratio between the average simulation standard error and standard deviation over replications (“s.e./sd”).

It can be seen in Table 3 that under all conditions, the bias is small for most parameters and the estimated standard errors accurately reflect the simulation standard deviation. Exceptions to this good performance are the latent variable intercepts (e.g. α_{21} and κ_{11}) in the condition with the smallest sample size ($n = 200$). Although the bias in this condition is smaller for the other latent intercept parameters, there is a clear pattern of overestimating the size of the largest class and underestimating that of the other classes. This bias disappears as the sample size grows larger. The other parameters do not appear to show any bias, even at the smallest sample size.

Table 3 also shows the performance of information-based standard errors as an estimate of simulation standard deviation. The standard errors perform well when

sample size it at least 500. In the smallest sample size condition, some of the standard errors tend to underestimate the simulation standard deviation, which will lead to undercoverage of confidence intervals.

In summary, while the performance of the maximum-likelihood estimates is generally good, bias in some of the parameter estimates and many of the standard errors occurred when the sample size is small ($n = 200$). Therefore, we recommend to use the GMTMM model with samples of at least 500 cases.

5. DISCUSSION AND CONCLUSION

We showed how the quality of survey and administrative data can be evaluated using generalized multitrait-multimethod (GMTMM) models. This approach is an improvement over existing methods, which assume that either the survey or the administrative data are perfect measures. A general framework for data quality evaluation was introduced. This framework is more suited than existing MTMM approaches to administrative data particularities such as categorical measurement, nonlinearities, heterogeneous error processes, and nonnormality. We demonstrated the use of GMTMM models by applying them to administrative and survey data on income and duration of employment from the German Federal Employment Agency. A simulation study demonstrated good properties of the maximum-likelihood estimates for a GMTMM model with moderate sample sizes.

A clear advantage of our approach is that it allows for the presence of measurement error in both the survey and the administrative register. Furthermore, using the administrative register as a second measure in the MTMM design has an additional advantage over classical MTMM designs using repeated survey measures. When repeated survey measures are used, survey respondents must answer questions on the same topic twice and may remember their answer, creating depen-

dencies that are not modeled (Alwin, 2011), although van Meurs (1995) provided some evidence that this might not occur in practice when sufficient time is allowed between the repetitions. The problem of memory bias does not occur, however, when the measurement methods are administrative and survey data collected separately. Therefore, besides allowing for the estimation of measurement error in administrative records, the MTMM design using linked survey-register data is an attractive method of estimating measurement error in survey variables.

Some limitations of our work remain. First, we did not discuss model fit evaluation. However, this issue is not specific to GMTMM modeling, so that the standard machinery available for global and local fit assessment in generalized latent variable models can trivially be applied to GMTMM modeling (see, e.g. Skrondal and Rabe-Hesketh, 2004; Oberski and Vermunt, 2013; Oberski et al., 2013). Second, little is known about the small sample properties of GMTMM model estimates. While simulation results by Scholtus and Bakker (2013) on the linear MTMM model were positive, other types of GMTMM models were not evaluated as to their stability and robustness. This remains a topic for future research. Finally, in our application on German data, unique identifiers were available that allowed for close linkage between the survey and register. In other applications, however, such identifiers may not be available for legal reasons or they may not exist. In such cases, linkage error will occur as well as measurement error. Incorporating such errors into the GMTMM model remains a topic for future study as well.

REFERENCES

- Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132.

- Alwin, D. F. (1973). Approaches to the interpretation of relationships in the multitrait-multimethod matrix. *Sociological methodology*, 5:79–105.
- Alwin, D. F. (2007). *Margins of error: a study of reliability in survey measurement*. Wiley-Interscience, New York.
- Alwin, D. F. (2011). Evaluating the reliability and validity of survey interview data using the MTMM approach. In Madans, J., Miller, K., Maitland, A., and Willis, G., editors, *Question Evaluation Methods: Contributing to the Science of Data Quality*, pages 263–293. Wiley Online Library, New York.
- Andrews, F. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, 48(2):409–442.
- Ansolabehere, S. and Hersh, E. (2010). The quality of voter registration records: A state-by-state analysis. *Cambridge, Mass.: Department of Government, Harvard University*.
- Ansolabehere, S. and Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*, 20(4):437–459.
- Bakker, B. F. (2009). *Trek alle registers open! [Open up the registers!]*. VU University, Amsterdam.
- Bakker, B. F. (2012). Estimating the validity of administrative variables. *Statistica Neerlandica*, 66(1):8–17.
- Biemer, P. (2011). *Latent Class Analysis of Survey Error*. Wiley, New York.
- Bollen, K. (1989). *Structural Equations with Latent Variables*. John Wiley & Sons, New York.

- Campbell, D. and Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56:81–105.
- Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2006). *Measurement error in nonlinear models: a modern perspective*, volume 105. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Boca Raton, FL.
- Catchpole, E. and Morgan, B. (1997). Detecting parameter redundancy. *Biometrika*, 84(1):187–196.
- Eckman, S., Kreuter, F., Kirchner, A., Jäckle, A., Tourangeau, R., and Presser, S. (2014). Assessing the mechanisms of misreporting to filter questions in surveys. *Public Opinion Quarterly*, 78(3):721–733.
- Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, 65:pp. 241–261.
- Entwisle, B. and Elias, P. (2013). *New Data for Understanding the Human Condition: International Perspectives*. OECD, Paris, France.
- Fellgi, I. (1964). Response variance and its estimation. *Journal of the American Statistical Association*, pages 1016–1041.
- Forcina, A. (2008). Identifiability of extended latent class models with individual covariates. *Computational Statistics & Data Analysis*, 52(12):5263–5268.
- Fuller, W. (1987). *Measurement Error Models*. John Wiley & Sons, New York.
- Gerardi, K., Goette, L., and Meier, S. (2013). Numerical ability predicts mortgage default. *Proceedings of the National Academy of Sciences*, 110(28):11267–11271.

- Gottschalk, P. and Huynh, M. (2010). Are earnings inequality and mobility overstated? the impact of nonclassical measurement error. *The Review of Economics and Statistics*, 92(2):302–315.
- Groen, J. A. (2012). Sources of error in survey and administrative data: The importance of reporting procedures. *Journal of Official Statistics (JOS)*, 28(2).
- Hansen, M., Hurwitz, W., and Bershad, M. (1961). Measurement errors in censuses and surveys. *Bulletin of the International Statistical Institute*, 38(2):359–374.
- Hansen, M., Hurwitz, W., and Pritzker, L. (1964). The estimation and interpretation of gross differences and the simple response variance. In Rao, C. R., editor, *Contributions to Statistics, Presented to Professor P. C. Mahalanobis on the Occasion of his 70th Birthday*. Pergamon Press, Calcutta, India.
- Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences*. Sage, Thousand Oaks, CA.
- Japac, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O’Neil, C., and Usher, A. (2015). *AAPOR Report on Big Data*. American Association for Public Opinion Research (AAPOR).
- Kapteyn, A. and Ypma, J. Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, 25(3):513–551.
- Kim, C. and Tamborini, C. R. (2014). Response error in earnings an analysis of the survey of income and program participation matched with administrative data. *Sociological Methods & Research*, 43(1):39–72.

- Kreuter, F., Müller, G., and Trappmann, M. (2010). Nonresponse and measurement error in employment research nonresponse and measurement error in employment research: Making use of administrative data. *Public Opinion Quarterly*, 75(5):pp. 880–906.
- Ladouceur, M., Rahme, E., Pineau, C. A., and Joseph, L. (2007). Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics*, 63(1):272–279.
- Lauritzen, S. L. (1996). *Graphical models*. Oxford University Press, Oxford, UK.
- Lord, F. M. and Novick, M. R. (1968). *Statistical theories of mental scores*. Addison–Wesley, Reading.
- McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons, New York.
- McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley, New York.
- Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, 22:43–65.
- Nordberg, L., Rendtel, U., and Basic, E. (2004). Measurement error of survey and register income. *Harmonisation of Panel Surveys and Data Quality*, pages 65–88.
- Oberski, D. (2013). The latent class MTMM model. *Psychological Methods*.
- Oberski, D., Saris, W. E., and Hagenaars, J. (2008). Categorization errors and differences in the quality of questions across countries. In Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., Pennell, B.-E., and Smith, T. W., editors, *Survey Methods in Multinational, Multiregional, and Multicultural Contexts (3MC)*. Wiley, New York.

- Oberski, D., Van Kollenburg, G., and Vermunt, J. (2013). A Monte Carlo evaluation of three methods to detect local dependence in binary data latent class models. *Advances in Data Analysis and Classification*, 7(3).
- Oberski, D. and Vermunt, J. (2013). A model-based approach to goodness-of-fit evaluation in item response theory. *Measurement: Interdisciplinary Research & Perspectives*, 11:117–122.
- Pavlopoulos, D. and Vermunt, J. (2013). Measuring temporary employment. do survey or register data tell the truth? *Survey Methodology*.
- Podesta, J. (2014). *Big Data: Seizing Opportunities Preserving Values*. Executive Office of the President.
- R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
- Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, 69(2):167–190.
- Rindfuss, R. R., Walsh, S. J., Turner, B., Fox, J., and Mishra, V. (2004). Developing a science of land change: challenges and methodological issues. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39):13976–13981.
- Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., and Savalei, V. (2013). *lavaan*: Latent variable analysis. [Software]. Available from <http://lavaan.ugent.be/>.
- Sakshaug, J. W., Yan, T., and Tourangeau, R. (2010). Nonresponse error, measurement error, and mode of data collection: tradeoffs in a multi-mode survey of sensitive and non-sensitive items. *Public Opinion Quarterly*, 74(5):907–933.

- Samejima, F. (1969). *Estimation of latent trait ability using a response pattern of graded scores*, volume 17 of *Psychometrika Monograph*. Psychometric Society, Bowling Green, OH.
- Saris, W. E. and Andrews, F. M. (1991). Evaluation of measurement instruments using a structural modeling approach. In Biemer, P., Groves, R., Lyberg, L., Mathiowetz, N., and Sudman, S., editors, *Measurement errors in surveys*, pages 575–599. John Wiley & Sons, New York.
- Saris, W. E. and Gallhofer, I. N. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Wiley-Interscience, New York.
- Scholtus, S. (2015). validiteit en vertekening van administratieve data [validity and bias in administrative data]. In *Innovatie in survey onderzoek [innovation in survey research]*, Statistics Netherlands, Den Haag. NPSO [Dutch-language platform for Survey Research].
- Scholtus, S., Bakker, B., and van Delden, A. (2015). Validiteit en vertekening van administratieve data [validity and bias in administrative data]. Statistics Netherlands. Dutch-language platform for survey research (NPSO). <http://www.npsso.net/sites/default/files/2-3%20Scholtus%20-%20voor%20website.pdf>.
- Scholtus, S. and Bakker, B. F. (2013). *Estimating the validity of administrative and survey variables through structural equation modeling: a simulation study on robustness*. CBS Discussion Papers. Statistics Netherlands, The Hague.
- Shapiro, A. and Browne, M. (1983). On the investigation of local identifiability: A counterexample. *Psychometrika*, 48(2).
- Skrondal, A. and Kuha, J. (2012). Improved regression calibration. *Psychometrika*, 77(4):649–669.

- Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling : multilevel, longitudinal, and structural equation models*. Interdisciplinary statistics series. Chapman & Hall/CRC, Boca Raton, FL.
- Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica: journal of the Econometric Society*, pages 24–36.
- van Meurs, L. (1995). Memory effects in MTMM studies. In Saris, W. E. and Münnich, A., editors, *The multitrait-multimethod approach to evaluate measurement instruments*. Eötvös University Press, Budapest.
- Vermunt, J. (2010). Latent class modeling with covariates: Two improved three-step approaches. *Political Analysis*, 18:450–469.
- Vermunt, J. K. and Magidson, J. (2004). Factor analysis with categorical indicators: A comparison between traditional and latent class approaches. In van der Ark, L. A., Croon, M. A., and Sijtsma, K., editors, *New developments in categorical data analysis for the social and behavioral sciences*, pages 41–63. Erlbaum, Mahwah.
- Vermunt, J. K. and Magidson, J. (2013). *Technical guide for Latent GOLD 5.0: Basic, Advanced, and Syntax*. Statistical Innovations Inc., Belmont, MA.
- Wallgren, A. and Wallgren, B. (2007). *Register-based statistics: administrative data for statistical purposes*. Wiley, New York.
- Yucel, R. M. and Zaslavsky, A. M. (2005). Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association*, 100(472):1123–1132.

Table 3: Simulation results for a generalized MTMM model, under different sample sizes. Shown are the true values of the parameters, the simulation bias, and the ratio between the average simulation standard error and standard deviation over replications (“s.e./sd”).

Par.	True	Sample size n							
		200		500		1000		2000	
		Bias	s.e./sd	Bias	s.e./sd	Bias	s.e./sd	Bias	s.e./sd
α_{11}	0.889	0.013	0.956	-0.001	1.002	-0.002	0.968	-0.002	1.013
α_{12}	0.085	-0.009	1.001	0.004	1.088	0.008	1.067	0.004	0.994
α_{21}	1.426	0.074	0.875	0.027	0.964	0.015	0.962	0.013	0.965
α_{22}	-0.305	-0.013	0.943	-0.002	0.999	-0.010	1.020	-0.006	0.985
α_{31}	-0.121	0.017	0.996	-0.003	1.040	-0.007	0.960	-0.002	0.955
α_{32}	-0.356	-0.007	0.948	0.008	1.015	0.010	1.021	0.006	1.069
κ_{11}	0.058	0.018	0.752	0.005	0.902	0.005	0.920	0.001	0.939
κ_{21}	-0.888	-0.015	0.917	-0.008	0.967	-0.003	0.940	-0.005	1.001
τ_{11}	1.296	0.001	0.940	0.003	0.963	-0.000	1.042	-0.001	1.013
λ_{11}	3.772	-0.017	0.815	-0.004	0.917	-0.000	0.948	0.007	0.943
γ_{11}	-1.025	-0.007	1.047	-0.003	1.022	-0.004	1.105	-0.002	0.983
τ_{21}	0.693	-0.015	0.943	-0.000	1.049	0.004	1.065	0.003	1.096
λ_{21}	1.546	0.013	0.956	-0.001	1.005	-0.005	1.010	0.002	0.998
γ_{11}	0.043	0.031	0.850	0.008	0.953	-0.000	0.973	-0.003	0.954
τ_{31}	0.366	0.001	0.870	0.000	0.988	-0.000	0.943	-0.000	0.991
λ_{31}	-0.283	-0.001	0.931	-0.000	1.090	0.000	1.032	0.000	1.008
γ_{31}	0.001	-0.001	0.830	-0.001	0.961	-0.000	1.050	-0.000	1.061
τ_{12}	4.811	0.004	1.025	0.000	1.015	0.005	1.014	0.004	0.950
λ_{12}	2.029	0.003	0.929	-0.001	0.988	-0.004	0.992	-0.003	0.987
γ_{12}	-3.169	-0.003	1.026	0.002	1.023	-0.001	1.038	-0.002	0.958
τ_{22}	1.017	0.009	0.915	0.002	0.982	-0.001	0.947	0.002	0.968
λ_{22}	1.964	-0.003	0.981	-0.001	1.020	0.001	0.960	0.002	0.970
γ_{22}	-0.224	-0.002	0.902	0.001	1.019	0.003	0.966	-0.000	0.967
τ_{32}	0.384	0.001	0.959	-0.000	0.945	0.000	0.968	0.001	1.094
λ_{32}	-0.114	-0.002	0.971	-0.000	0.943	-0.000	0.961	-0.001	0.998
γ_{32}	-0.006	-0.001	0.963	-0.001	0.995	-0.000	1.006	-0.001	1.099
ϕ_{12}	2.916	0.067	0.882	0.032	1.001	0.020	0.969	0.009	0.986
ϕ_{13}	-0.992	-0.012	0.895	-0.033	0.950	-0.008	0.912	-0.000	0.997
ϕ_{23}	-0.289	0.059	0.872	0.020	0.986	0.005	1.016	0.012	0.998
$\sigma_{\epsilon,11}$	0.175	0.004	0.771	0.001	0.934	-0.001	1.005	-0.001	0.984
$\sigma_{\epsilon,21}$	0.420	-0.017	0.993	-0.007	0.971	-0.004	1.055	-0.003	1.074
$\sigma_{\epsilon,31}$	0.003	-0.000	0.891	-0.000	1.031	-0.000	0.932	-0.000	0.941
$\sigma_{\epsilon,12}$	0.545	-0.005	1.043	-0.005	0.931	-0.002	0.940	-0.002	0.980
$\sigma_{\epsilon,22}$	0.141	-0.002	1.067	0.001	1.043	-0.000	1.064	0.000	0.954
$\sigma_{\epsilon,32}$	0.015	-0.000	1.030	-0.000	0.993	-0.000	1.039	-0.000	1.081